Open-Source Toolkit for Simple XML Annotation

AMIA 2003 Open Source Expo

Burke W. Mamlin, M.D., Gunther Schadow, M.D., Ph.D., J. Marc Overhage, M.D.
Regenstrief Institute for Health Care, Indianapolis, Indiana

ABSTRACT

Use of Extensible Markup Language (XML) is increasingly prevalent among medical informatics projects. Many of these projects involve, at some point, the interaction between a researcher and specialized XML documents for the purpose of annotating the XML data. We offer a simple toolkit to assist these researchers. Our solution is a simple, yet fully functional, annotation system that can easily be adapted to the needs of the researcher. All of the materials for this toolkit are freely available.

THE PROBLEM

We have recently encountered several instances where researchers were working with XML documents and applying various forms of annotation to the documents. In one case, we marked up dictated reports from a speech recognition engine in order to create an error classification scheme.[1] In two other cases, we marked up XML documents to evaluate specific language parsing algorithms.[2, 3] In each case, XSLT was employed to render a more visually accommodating form of the data. * Across these projects we wanted to work with rendered (post XSLT-transformation) data, while applying the effect of our interaction to the original XML document. While existing applications can help in many cases, we believe that a simplified, well-documented, working toolkit would better suit the individual needs of researchers.

THE SOLUTION

We developed a simple method of annotating the XML documents, using a combination of a Perl- or Java-based server routine, XSLT, and some JavaScript (see Figure 1).

In the spirit of open source, we would like to share our "poor man's XML annotation toolkit" as a resource for other informatics researchers facing a similar situation. The toolkit may be obtained by contacting the authors by e-mail (bmamlin@regenstrief.org).

ACKNOWLEDGEMENTS

This research was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was funded, in part, by National Cancer Institute grant U01 CA91343, a Cooperative Agreement for The Shared Pathology Informatics Network, and by National Library of Medicine grant T15 LM07117-06.

REFERENCES

1. Zafar A, Mamlin B, Overhage JM, Belsito A, McDonald CJ. A Semantic Classification System for Converting Speech to Text. JAMIA (in review).
2. Mamlin BW, Heinze DT, McDonald CJ. Automated Extraction and Normalization of Findings from Free-Text Radiology Reports. Proceeding of the AMIA Annual Symposium (in review).
3. Schadow G, McDonald CJ. Parsing Structured Information from Free Text Pathology Reports. Proceeding of the AMIA Annual Symposium (in review).

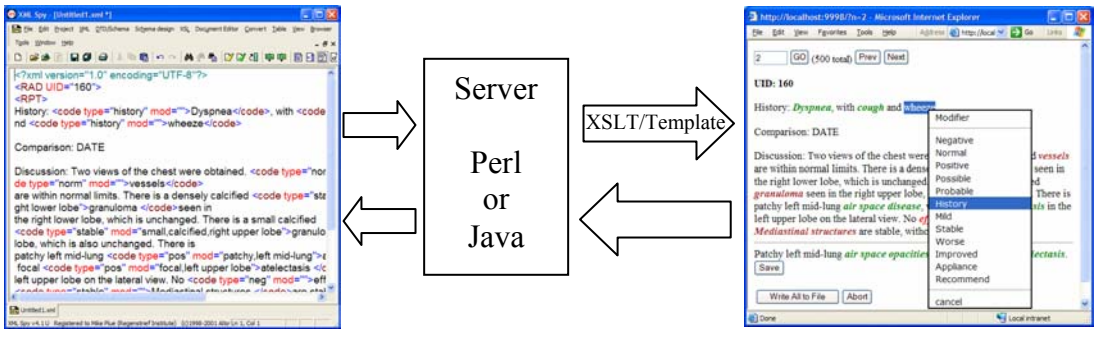

Figure 1. A Sample Implementation of the XML Annotation Toolkit. XML data is read by Perl or Java program acting as a server. The browser renders a visual representation of the XML (using XSLT or an HTML template) and annotations are made using JavaScript for convenience. Annotations are then submitted back to the server program and the XML data is updated accordingly.

* In this case a Visual Basic® application was created to assist the researchers